Exploring the Landscape of Spatial Robustness

Logan Engstrom

(with Brandon Tran*, Dimitris Tsipras*, Ludwig Schmidt, Aleksander Mądry)

madry-lab.ml

ML “Glitch”: Adversarial Examples

ML “Glitch”: Adversarial Examples

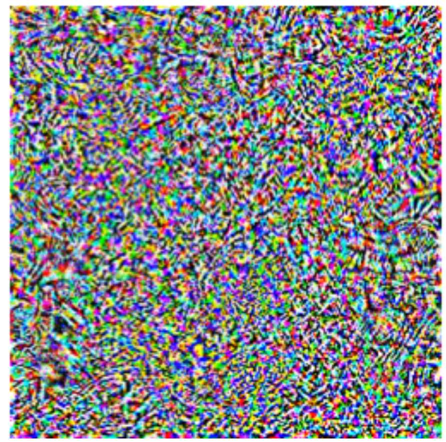
“pig”

ML “Glitch”: Adversarial Examples

“pig”

+ 0.005 x

small, *non-random* noise

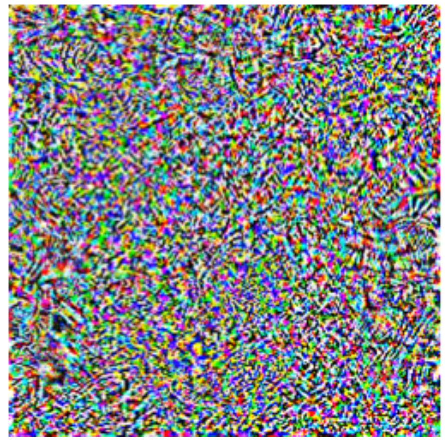


ML “Glitch”: Adversarial Examples

“pig”

+ 0.005 x

small, *non-random* noise



=

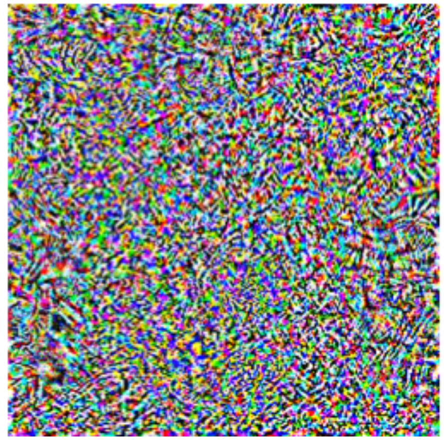
“airliner”

ML “Glitch”: Adversarial Examples

“pig”

+ 0.005 x

small, *non-random* noise



=

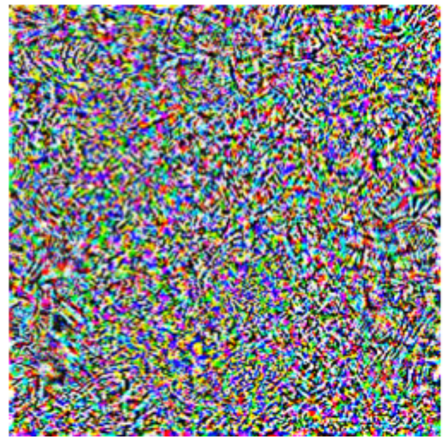
“airliner”

What does **small** mean here?

ML "Glitch": Adversarial Examples

"pig" **small**, *non-random* noise "airliner"

+ 0.005 x



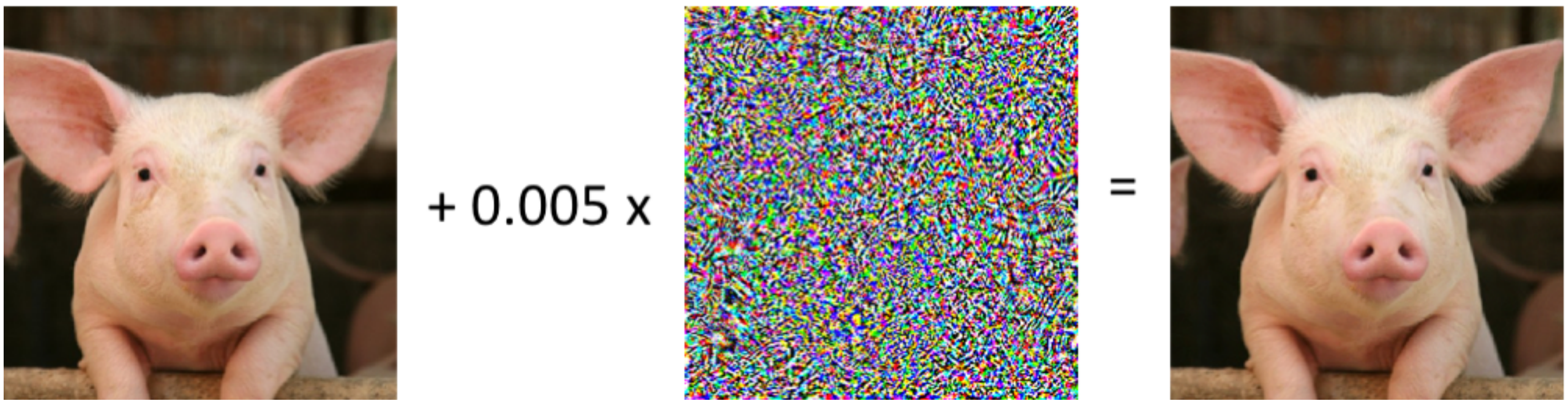
=

What does **small** mean here?

Traditionally: perturbations that have small **l_p norm**

ML “Glitch”: Adversarial Examples

“pig” **small**, *non-random* noise “airliner”



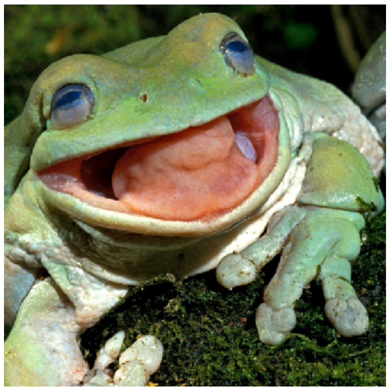
What does **small** mean here?

Traditionally: perturbations that have small **l_p norm**

Do small l_p norms capture every sense of “small”?

Spatial Perturbations

Spatial Perturbations



Spatial Perturbations

rotation up to 30°

Spatial Perturbations

rotation up to 30°　　　x, y translations up to ~10%

Spatial Perturbations

rotation up to 30°   x, y translations up to ~10%

These are **not** small l_p perturbations!

Spatial Perturbations

rotation up to 30°   x, y translations up to ~10%

These are **not** small l_p perturbations!

How robust are models to spatial perturbations?

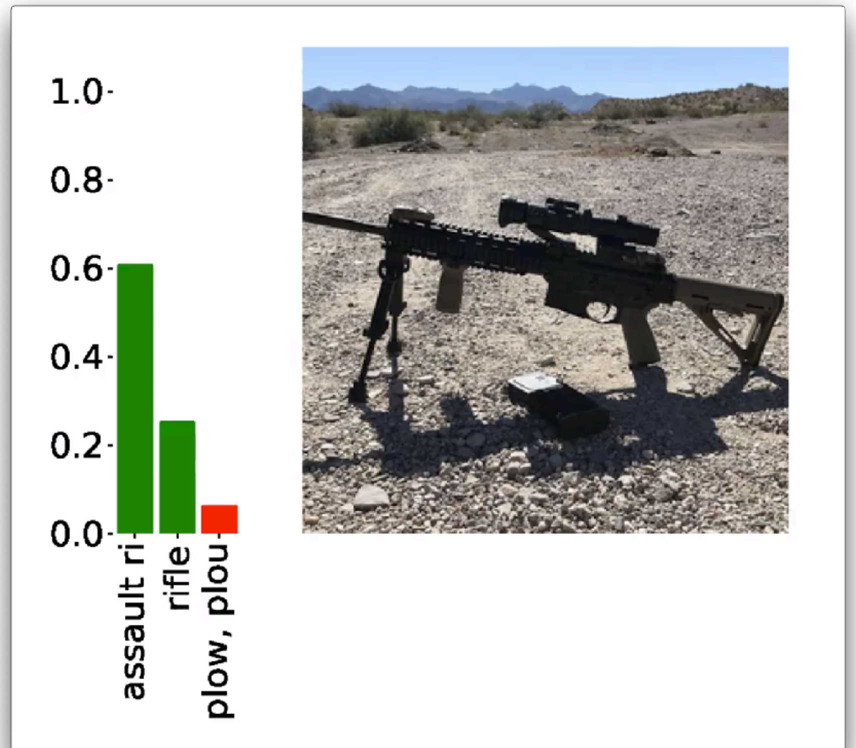
Spatial Robustness

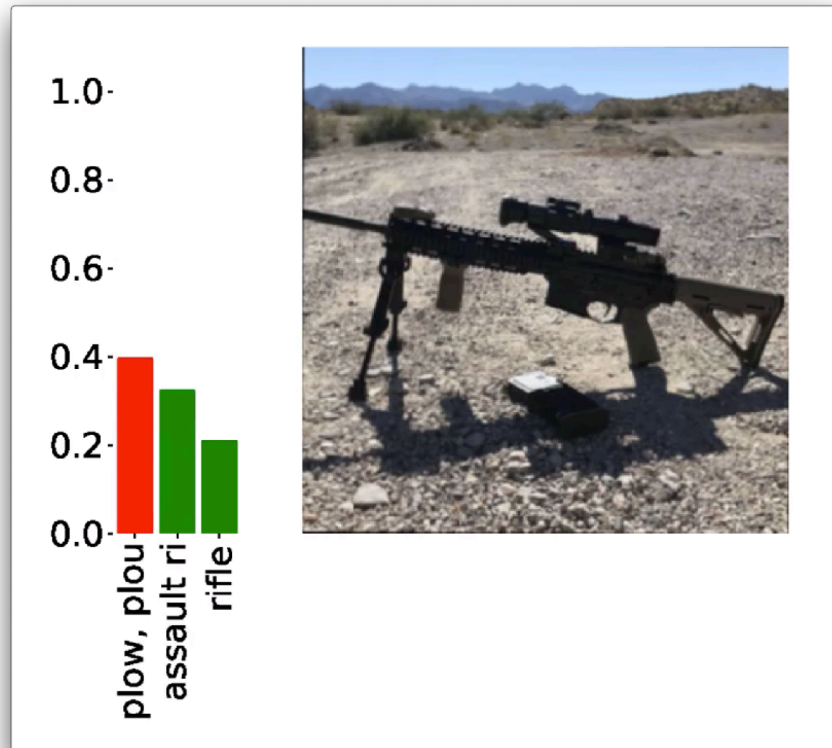
Spatial Robustness

Spoiler: models are not robust

Spatial Robustness

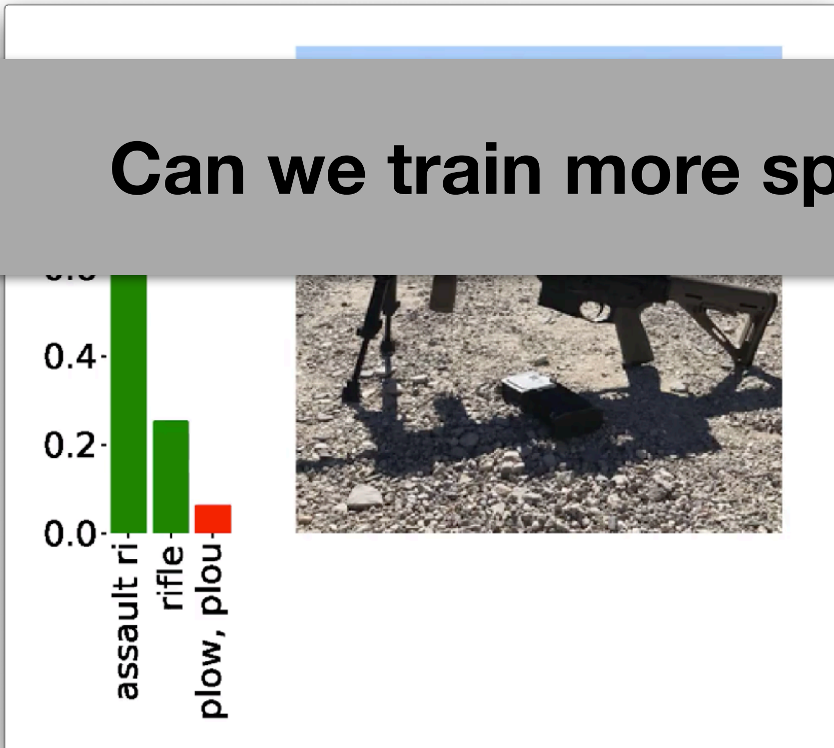
Spoiler: models are not robust

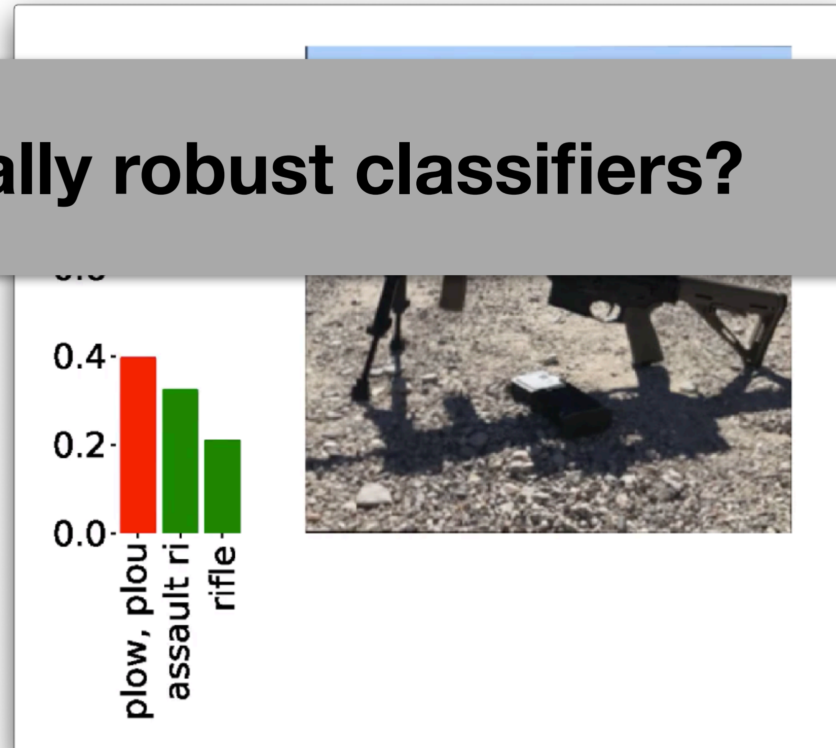




Spatial Robustness

Spoiler: models are not robust





Can we train more spatially robust classifiers?

Spatial Defenses

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**

(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

Spatial Defenses

Lesson from l_p robustness: use **robust optimization** (= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

Key question: how to find **worst-case** translations, rotations?

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al '15][Madry et al '18]

Key question: how to find **worst-case** translations, rotations?

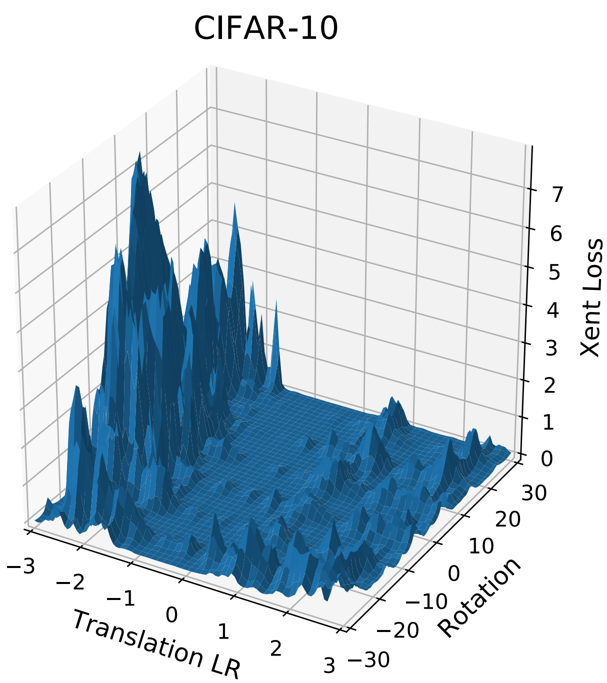
Attempt #1: first-order methods

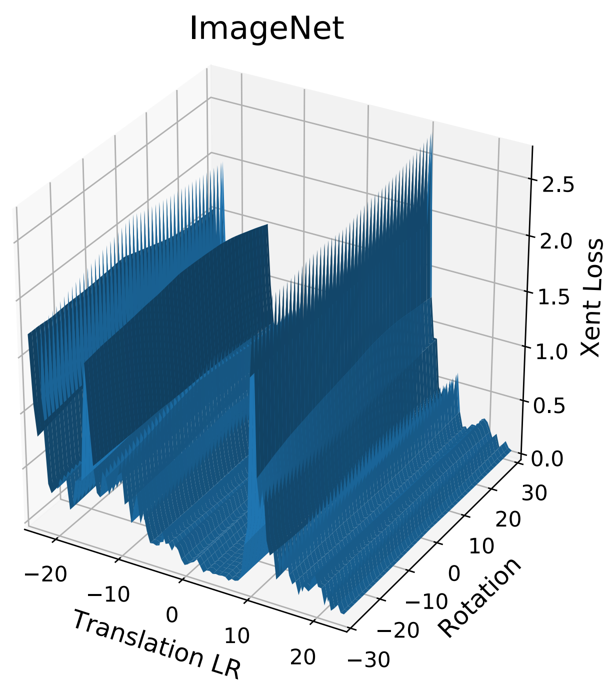
Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al '15][Madry et al '18]

Key question: how to find **worst-case** translations, rotations?

~~Attempt #1: first order methods~~



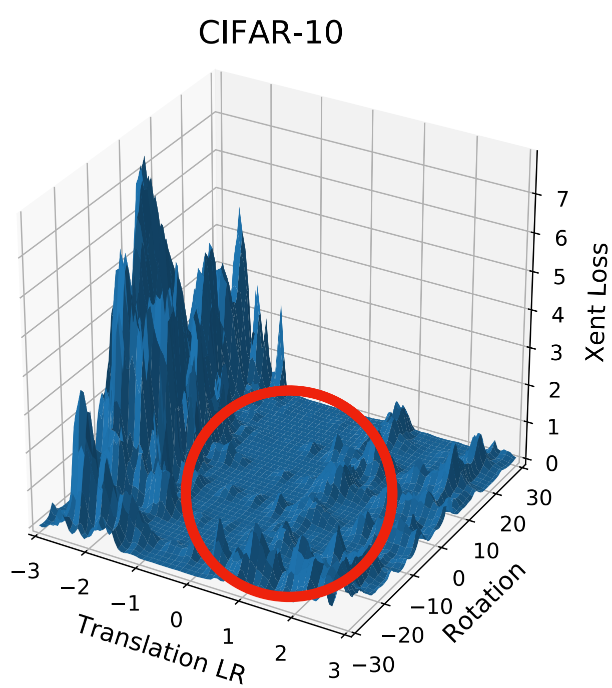


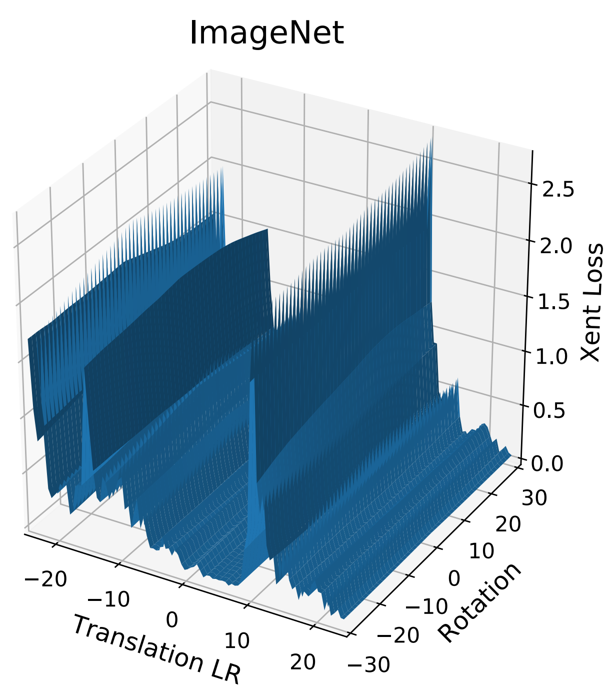
Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

Key question: how to find **worst-case** translations, rotations?

~~Attempt #1: first order methods~~





Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

Key question: how to find **worst-case** translations, rotations?

~~Attempt #1: first-order methods~~

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al '15][Madry et al '18]

Key question: how to find **worst-case** translations, rotations?

~~**Attempt #1: first order methods**~~

Attempt #2: exhaustive search

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al '15][Madry et al '18]

Key question: how to find **worst-case** translations, rotations?

~~**Attempt #1: first order methods**~~

Attempt #2: exhaustive search

Exhaustive search is feasible, and a strong adversary!

(discretize translations and rotations, try every combination)

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

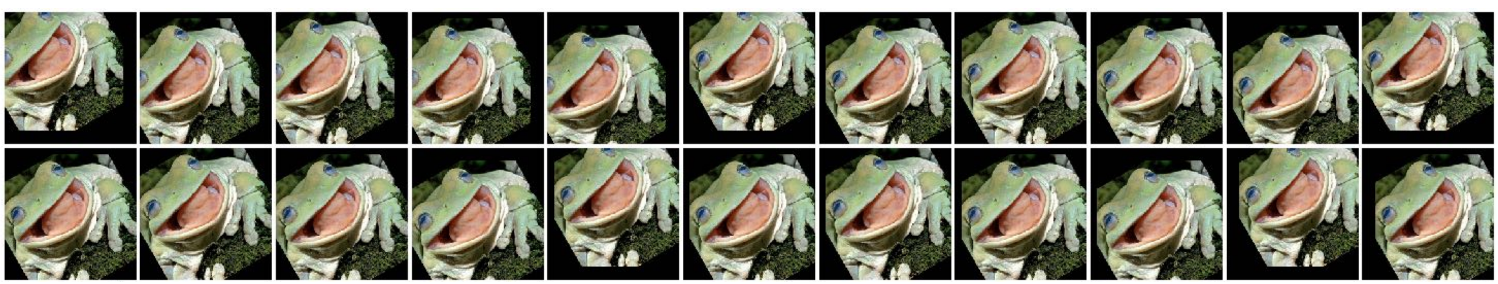
Key question: how to find **worst-case** translations, rotations?

~~**Attempt #1: first order methods**~~

Attempt #2: exhaustive search

Exhaustive search is feasible, and a strong adversary!

(discretize translations and rotations, try every combination)



Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

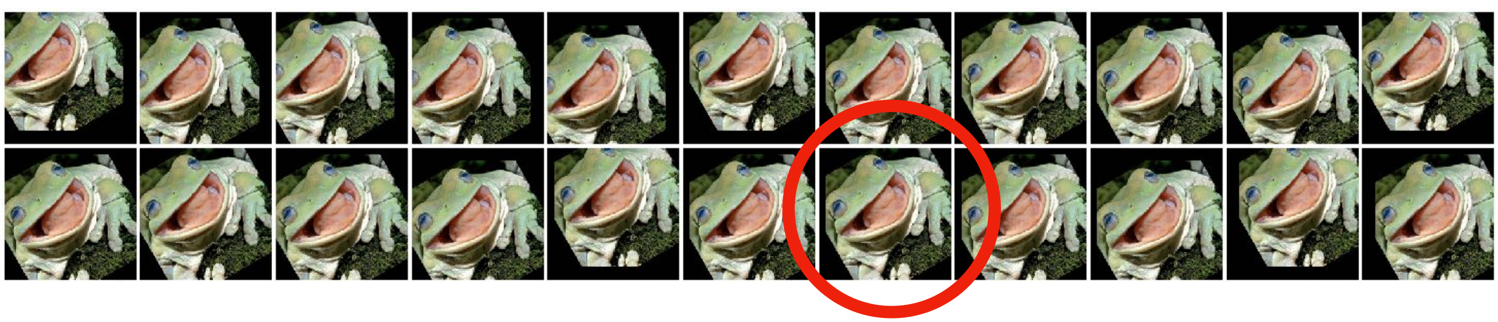
Key question: how to find **worst-case** translations, rotations?

~~**Attempt #1: first order methods**~~

Attempt #2: exhaustive search

Exhaustive search is feasible, and a strong adversary!

(discretize translations and rotations, try every combination)



Train only on “worst” transformed input (highest loss)

Spatial Defenses

Lesson from l_p robustness: use **robust optimization**
(= train on **worst-case** perturbed inputs) [Goodfellow et al ‘15][Madry et al ’18]

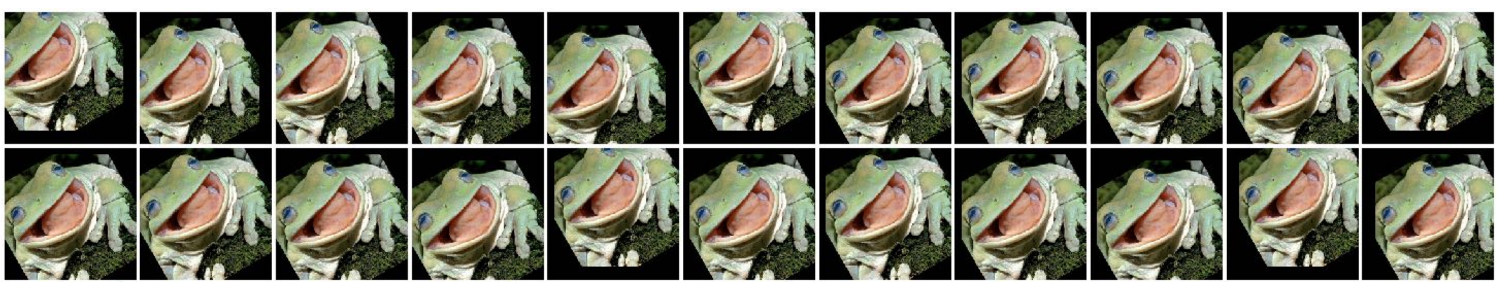
Key question: how to find **worst-case** translations, rotations?

~~**Attempt #1: first-order methods**~~

Attempt #2: exhaustive search

Exhaustive search is feasible, and a strong adversary!

(discretize translations and rotations, try every combination)



(we approximate via 10 random samples to quicken training)

Spatial Defenses

With robust optimization:

Spatial Defenses

With robust optimization:

CIFAR classifier accuracy: 3% adversarial to **71% adversarial**

Spatial Defenses

With robust optimization:

CIFAR classifier accuracy: 3% adversarial to **71% adversarial**
(compare to **93%** standard accuracy)

Spatial Defenses

With robust optimization:

CIFAR classifier accuracy: 3% adversarial to **71% adversarial**
(compare to **93%** standard accuracy)

ImageNet classifier accuracy: 31% adversarial to **53% adversarial**

Spatial Defenses

With robust optimization:

CIFAR classifier accuracy: 3% adversarial to **71% adversarial**
(compare to **93%** standard accuracy)

ImageNet classifier accuracy: 31% adversarial to **53% adversarial**
(compare to **76%** standard accuracy)

Spatial Defenses

With robust optimization:

(+10 sample majority vote)

CIFAR classifier accuracy: 3% adversarial to **71% adversarial**

(compare to **93%** standard accuracy)

ImageNet classifier accuracy: 31% adversarial to **53% adversarial**

(compare to **76%** standard accuracy)

Spatial Defenses

With robust optimization:

(+10 sample majority vote) 82%

CIFAR classifier accuracy: 3% adversarial to **~~71~~% adversarial**

(compare to **93%** standard accuracy)

ImageNet classifier accuracy: 31% adversarial to **53% adversarial**

(compare to **76%** standard accuracy)

Spatial Defenses

With robust optimization:

(+10 sample majority vote)

82%

CIFAR classifier accuracy: 3% adversarial to **~~71~~% adversarial**

(compare to **93%** standard accuracy)

56%

ImageNet classifier accuracy: 31% adversarial to **~~53~~% adversarial**

(compare to **76%** standard accuracy)

Spatial Defenses

With robust optimization:

(+10 sample majority vote) 82%

CIFAR classifier accuracy: 3% adversarial to **~~71~~% adversarial**

(compare to **93%** standard accuracy)

56%

ImageNet classifier accuracy: 31% adversarial to **~~58~~% adversarial**

(compare to **76%** standard accuracy)

Still significant room for improvement!

Conclusions

Conclusions

Robust models need more refined notions of similarity

Conclusions

Robust models need more refined notions of similarity

We do not have true spatial robustness

Conclusions

Robust models need more refined notions of similarity

We do not have true spatial robustness

Intuitions from l_p robustness do not transfer

Conclusions

Robust models need more refined notions of similarity

We do not have true spatial robustness

Intuitions from l_p robustness do not transfer

Come to our poster! Pacific Ballroom #142